Can Anthropographics Promote Prosociality? A Review and Large-Sample Study

Luiz Morais
Université Paris-Saclay, CNRS, Inria, LISN
Gif-sur-Yvette, France
luiz.morais@inria.fr

Yvonne Jansen
Sorbonne Université, CNRS, ISIR
Paris, France
yvonne.jansen@sorbonne-universite.fr

Nazareno Andrade
Universidade Federal de Campina Grande
Campina Grande, Brazil
nazareno@computacao.ufcg.edu.br

Pierre Dragicevic
Université Paris-Saclay, CNRS, Inria, LISN
Gif-sur-Yvette, France
pierre.dragicevic@inria.fr

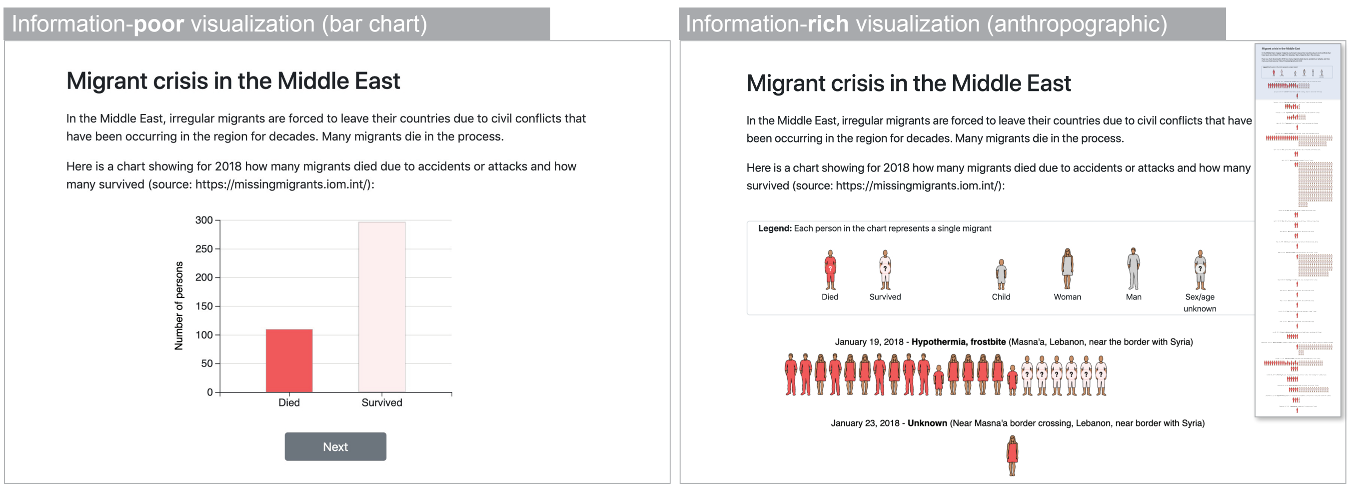

Figure 1: Two stimuli used in experiment 1, showing data from the Missing Migrants Project (missingmigrants.iom.int). On the left is the information-poor baseline condition, and on the right is the information-rich anthropographic condition (participants had to scroll to see the entire visualization – see overview on the right).

ABSTRACT

Visualizations designed to make readers compassionate with the persons whose data is represented have been called anthropographics and are commonly employed by practitioners. Empirical studies have recently examined whether anthropographics indeed promote empathy, compassion, or the likelihood of prosocial behavior, but findings have been inconclusive so far. This work contributes a detailed overview of past experiments, and two new experiments that use large samples and a combination of design strategies to maximize the possibility of finding an effect. We tested an information-rich anthropographic against a simple bar chart, asking participants to allocate hypothetical money in a crowdsourcing study. We found that the anthropographic had, at best, a small effect on money allocation. Such a small effect may be relevant for large-scale donation campaigns, but the large sample sizes required to observe an effect and the noise involved in measuring it make it very difficult to study in more depth. Data and code are available at https://osf.io/xqae2/.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **User studies**; **Empirical studies in HCI**.

KEYWORDS

information visualization, infographics, anthropographics, empathy, compassion.

CHI '21, May 8–13, 2021, Yokohama, Japan

The definitive Version of Record was published in *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan*, https://doi.org/10.1145/3411764.3445637.

ACM Reference Format:
Luiz Morais, Yvonne Jansen, Nazareno Andrade, and Pierre Dragicevic. 2021. Can Anthropographics Promote Prosociality? A Review and Large-Sample Study. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3411764.3445637

1 INTRODUCTION

Designing visualizations[1] with the intent to make readers relate to the persons behind the data has become a common practice among data journalists and visualization designers. Examples include infographics of gun victims, the plight of refugees, or COVID-19 deaths. Such visualizations have been termed *anthropographics* [10, 42]. Design strategies for creating anthropographics are diverse, as can be seen in the corpora[2] curated by Boy et al. [10] and Morais et al. [42]. These examples bear witness of practitioners' intuitions for telling stories about individuals through their data, with the intent to make readers more empathetic and compassionate, and perhaps inspire them to act.

Over the past few years, researchers have started to empirically test these intuitions [10, 13, 15, 38, 43]. Although many studies so far have yielded inconclusive results, the design space of anthropographics is vast [42], with many possible design choices, and only some points in this space have so far been tested. In addition, many studies had relatively low statistical power, meaning that they could have missed a small or a medium effect. More studies are needed to understand if anthropographics can promote prosocial feelings or behavior, and if so, to what extent.

This paper offers two major contributions. First, it provides the first detailed overview of past studies on anthropographics. Second, it reports on two novel experiments. Our review of past studies revealed that all the anthropographics tested so far conveyed little or no authentic information about individuals. Yet it is possible that doing so can help people relate, consistent with research in psychology suggesting that so-called "statistical victims" are less likely to be on the receiving end of compassion than "identifiable victims" [49]. Some of the prior anthropographics studies did attempt to make the individuals appear more unique by adding *imaginary* details about them (e.g., silhouettes, names), but this strategy was not found to be beneficial [10, 15, 38]. We therefore report on a study that tests a class of anthropographic designs we refer to as *information-rich*, and which shows actual data about individuals (see Figure 1-right and Figure 5).

In addition to using a yet-untested design strategy, our study attempts to address some issues we identified in prior work. First, it addresses the issue of statistical power through larger samples, as well as experimental manipulations and measurements designed to maximize effect size and minimize noise (one of them was to combine multiple experimental manipulations leading to two very different visualization designs, see Figure 1). In addition, some of the prior work potentially suffers from issues of multiplicity [37] and analytical flexibility [54] that weaken the strength of evidence in their findings. We address this by pre-registering our data collection, exclusion and analysis procedures, and by identifying a single primary outcome per experiment.

Like much of previous work, our first experiment yielded inconclusive results. Our second experiment, which used a fairly large sample size (N = 788) and further improvements in measurement, yielded some evidence for a small effect of the anthropographic design on average money allocation (Cohen's d = 0.14, 95% CI [-0.001, 0.28]), our primary outcome. We also found strong evidence that the anthropographic caused people to report slightly more negative emotions.

One of our initial goals was to untangle the effect of different anthropographic design strategies, but we did not proceed further due to the effect being small and noisy and thus hard to measure. It stands to reason that less drastic and "fairer" comparisons than the tested bar chart against anthropographic design would lead to even smaller effect sizes. These can nonetheless be relevant in practice, such as providing a small increase in a large donation campaign. We discuss implications for our findings, such as the importance of testing alternative anthropographic design strategies as well as using alternative methodologies, such as real money donation tasks or observations in the field, for example, by testing alternative designs on the website of a real charity.

[1]We use "visualization" in a broad sense that encompasses communicative visualizations, i.e., infographics.
[2]See myjyby.github.io/Anthropographics/www and luizaugustomm.github.io/anthropographics/ for online catalogs.

2 BACKGROUND

In this section, we explain how the area of anthropographics intersects psychology and visualization. We also describe the terminology and design space on which we base the rest of this article.

2.1 Work from Psychology

Although there is no consensual definition of empathy or compassion in psychology, we refer here to empathy as *"the act of experiencing what you believe others are experiencing"* [7] and compassion as *"the feeling that arises in witnessing another's suffering and that motivates a subsequent desire to help"* [26]. While many essays and studies on anthropographics are concerned with empathy (e.g., [10, 15, 30]), compassion is more likely to lead to helping behavior [7, 39, 55]. Thus, although empathy is relevant to consider, whenever prosocial behavior is ultimately the desired response (e.g., donating money), compassion is at least as important.

Many studies on empathy and compassion involve charitable giving. Most charitable giving studies measure prosocial emotions (e.g., empathy or compassion) and prosocial behavior, such as donating or dedicating time for a cause [31]. For capturing emotions, studies usually ask participants to report how intensely they feel certain emotions on a Likert scale (e.g., sympathetic, compassionate, worried, sad, etc.) [3, 19]. Concerning prosocial behavior, some studies ask participants to donate hypothetical money (e.g., [19]), while others ask them to make real donations, typically by giving them money and asking them if they want to donate part of it to a real charity (e.g., [52]).

Research on charitable giving suggests that showing large-scale statistics of human tragedies often fails to evoke prosocial emotions or motivate action [51], which is known as the *psychic numbing effect*. A related effect is *compassion fade*, according to which the level of compassion towards people decreases as the number of suffering individuals increases [53]. Conversely, people tend to donate more and be compassionate with persons who can be identified by names or photographs [25], a bias that has been coined the *identifiable victim effect*. To counter compassion fade, it has been suggested to present numbers through visual narratives, focus on individuals instead of the large number, or tell an information-rich story of a particular victim [50].

2.2 Related Work in Visualization

Anthropographics relates to several research areas in visualization. *Visual embellishment* consists of embedding pictorial content into visualizations, which studies suggested can make charts more memorable [2, 8, 9] and engaging [27], though possibly also harder to process [8]. This area overlaps with anthropographics but it is more focused in terms of the design strategies it studies, and broader in the types of benefits it examines – anthropographics studies *any* design strategy that can promote *prosociality* specifically. There has been considerable research interest about *narrative visualization* and *storytelling* [34, 48], with multiple contributions in terms of frameworks, tools, techniques, and studies. This area overlaps with anthropographics in that it studies communicative visualizations. However, the scope of anthropographics is narrower due to its focus on visualizations that convey data about people and in techniques that are thought to promote prosocial feelings or behavior. Similarly, there has been work on how to convey very large numbers [14], some of which may be applied to convey large-scale suffering, but previous papers have not specifically focused on that goal. Finally, the area of *personal visualization* [29] relates to anthropographics due to its focus on conveying data about people, but the intent is rarely to instill empathy or compassion.

In the past few years, a number of web essays have been written by visualization practitioners discussing strategies to bring an audience closer to the persons whose data is shown (e.g., [28, 40, 47]). These give insights into the intuitions of designers but were not empirically tested. In academic research, Boy et al. [10] established the area of *anthropographics* by coining the term and conducting experiments (summarized in section 3). Follow-up work discussed novel types of designs, such as immersive charts [30] and data comics [1], but without empirical validation. A few studies have suggested that prosociality can be influenced by whether or not data is shown, and by the choice of data [21, 32, 46]. For example, people’s positive affect decreases after being exposed to data on genital mutilation [46], and they give more money to fight cancer when they are shown more worrisome data about cancer [21]. However, we are interested here in studies that examine the impact of visualization design, which we cover in detail in section 3.

2.3 Dimensions of Anthropographics

Comparing multiple anthropographic designs and experiments is easier with a design space in which it is possible to position and compare design strategies. Throughout this paper, we use the terminology, conceptual framework and design space of anthropographics introduced by Morais et al. [42], which is the most recent and thorough attempt to categorize such visualizations. Like them, we refer to anthropographics as:

> *visualizations that represent data about people in a way that is intended to promote prosocial feelings (e.g., compassion or empathy) or prosocial behavior (e.g., donating or helping).* [42]

The design space has seven dimensions. The dimensions are divided in two broad categories: *what* is shown and *how* it is shown. The next subsections summarize the definition of each dimension. Concrete examples and further motivation for using the dimensions can be found in Morais et al. [42].

2.3.1 What is shown. The four dimensions in this category capture “what information and how much information is presented in a visualization” [42].

- **Granularity** indicates how close the geometric marks in the visualization are to representing individual persons, as opposed to large groups of people. For example, a bar with data aggregated by country has low granularity, whereas a visualization that uses a pictogram for each person has maximum granularity.
- **Specificity** indicates in how far the attributes visualized results in distinctive marks. Continuing the previous pictogram example, if the pictograms convey no data about individuals and they all look identical, the visualization has low specificity. On the other hand, if each pictogram conveys rich and detailed information about the individual, the visualization has high specificity.
- **Coverage** indicates in how far the persons visualized cover the population featured in the visualization’s message. For example, a visualization about immigrants in a particular country could show data about all immigrants, or about a small subset of them who were interviewed.
- **Authenticity** indicates whether all information in the visualization is from an actual dataset or if synthetic information has been added. For example, using unique silhouettes for each person (as in Figure 2-bottom) or displaying their names results in partial authenticity unless these are the actual silhouettes or names of the persons visualized.

2.3.2 How it is shown. The three dimensions here capture “how information is represented” [42].

- **Realism** indicates how much the marks in a visualization resemble actual persons. For example, a visualization using photos of people as marks would be highly realistic whereas the use of simple dots would be low in realism.
- **Physicality** indicates in how far a visualization includes any physical components. For example, a visualization of unemployment shown on a website would be very low on physicality, whereas a data sculpture where each person is represented with a matchstick would be high in physicality.
- **Situatedness** indicates how physically close is the visualization to the people whose data is shown. Most web visualizations have low situatedness. Examples for high situatedness include jewelry of personal data, or a visualization of how many people passed by a certain bus stop, if the visualization is placed at exactly that bus stop. For the purposes of our study review, we extend the original definitions by Morais et al. by considering that situatedness can be temporal [56] and that it decreases with spatial or temporal distance [56].

This design space considers several dimensions, but it does not cover all possible design strategies. Nevertheless, we chose to frame our work according to Morais et al. [42]’s conceptual framework because it expands Boy et al. [10], which is the only other paper we know of that introduces a design space of anthropographics.

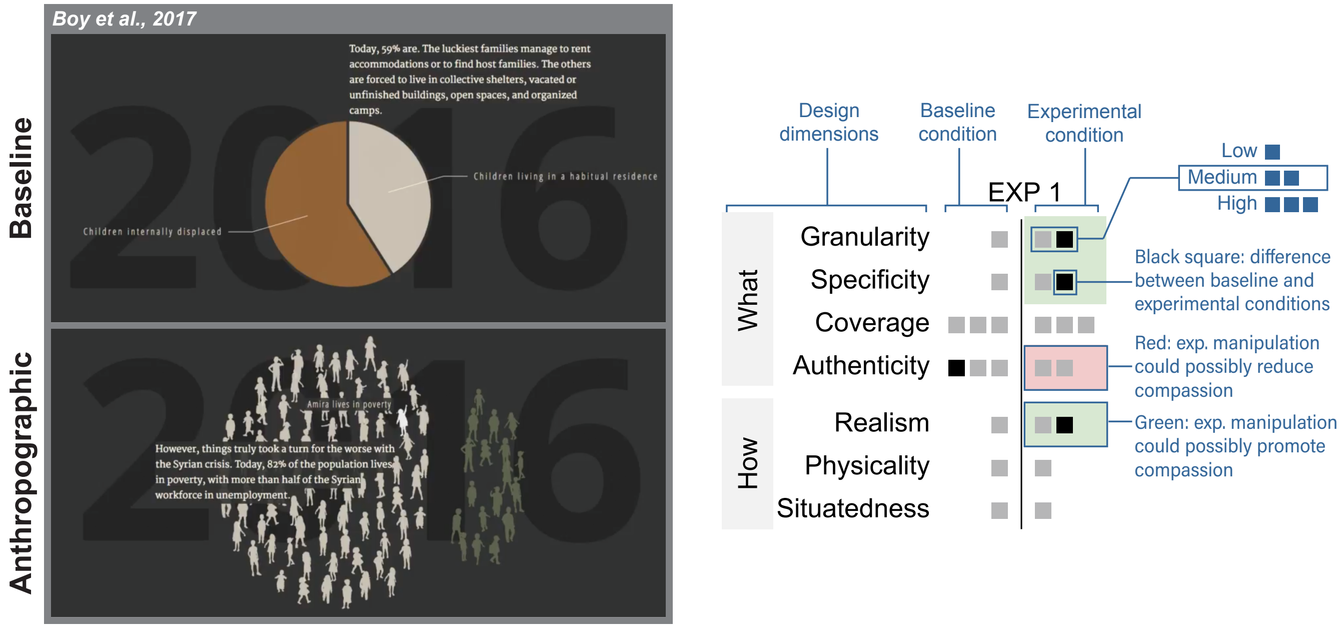

Figure 2: The left subfigure shows two stimuli from the first experiment of Boy et al. [10] (Credits: Jeremy Boy, Anshul Pandey, John Emerson, Margaret Satterthwaite, Oded Nov, and Enrico Bertini). The one at the top is from the pie chart condition, while the one at the bottom is from the anthropographic condition. The two charts also show different datasets. The right subfigure is a diagram summarizing the visualization designs involved in this experiment.

3 SUMMARY OF PREVIOUS STUDIES OF ANTHROPOGRAPHICS

A major difficulty when studying anthropographics is gathering and comparing information about previous studies (stimuli, measurements, findings), which is scattered across multiple articles, not presented consistently, and possibly hard to find (sometimes requiring examining supplementary material or reading a PhD dissertation). Thus we contribute an overview of empirical studies that have investigated the effect of visualization design on prosociality.

Boy et al. [10] coined the term anthropographics and were the first to study the effect of anthropographics on donation behavior. This summary includes work that cite Boy et al. and contribute a study testing anthropographic designs. We found three articles matching these criteria [13, 38, 43]. Additionally, we checked references in these three articles which resulted in one more article [15], and we performed a Google scholar search on the term "Anthropographics" which did not yield any new result. Note that this method was organic and not systematic, that is, we did not decide on a search methodology in advance.

One of the authors read the papers thoroughly, summarized their methods and findings, and categorized their stimuli. The study stimuli were coded according to the dimensions presented by Morais et al. [42], which provides extensive definitions for all dimensions and example visualizations. We did not use multiple coding with agreement analysis, but all authors were involved and reached a consensus for all stimuli. The studies are summarized in Figure 4.

3.1 Boy et al., 2017

The first and most comprehensive study is a crowdsourced study by Boy et al. [10]. In the first experiment, each participant saw two data-driven stories about the negative impact of the Syrian crisis on Syrian children. One story focused on poverty, while the other was on internally displaced people. Both were conveyed through an interactive slideshow [48] showing the proportion of affected people before the crisis and in the present time (2016). For one of the stories, this data was conveyed with a pie chart (see Figure 2-top-left), while for the other story, it was conveyed through an anthropographic design (Figure 2-bottom-left). The order was counterbalanced across participants.

The anthropographic designs conveyed proportions using 100 children silhouettes, each standing for 1% of the population. Each silhouette had a particular shape, and a first name that was displayed on mouse hover. Thus, this design differed from the baseline in four major ways: it had *medium granularity* (*low* for the baseline), *medium specificity* (vs. *low*), *partial authenticity* (vs. *full*), and *intermediate realism* (vs. *low*). The differences are summarized in the diagram of Figure 2-right, which we will use for visually summarizing all subsequent experiments. Each of the seven design dimensions from section 2.3 makes up one row on this diagram. The design of the baseline condition is summarized on the left of the vertical line, while the experimental condition is on the right. The number of squares (gray or black) indicate the value of the design on that dimension (low, medium, or high). Black squares highlight experimental manipulations (e.g., the experimental condition has an extra square on the *granularity* row, and one square less on the *authenticity* row). Finally, a green (resp. red) background

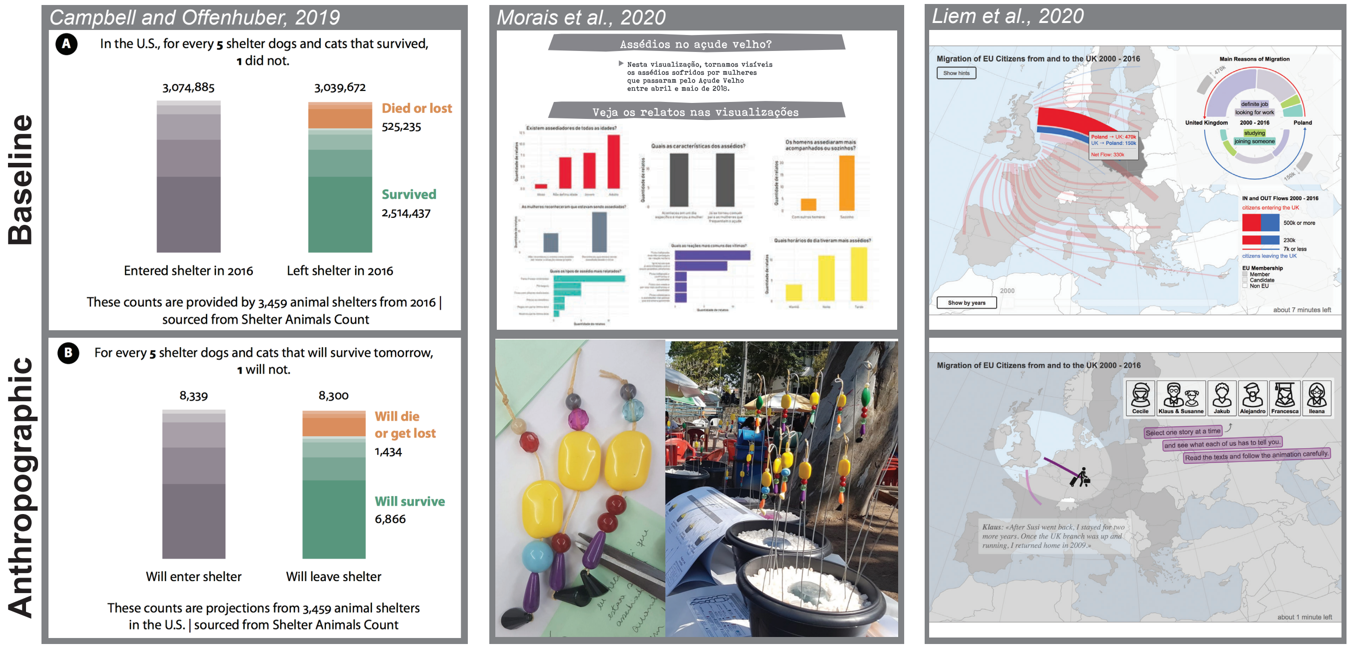

Figure 3: Stimuli used in the other three studies. Left: Campbell and Offenhuber [13] (Credits: Sarah Campbell and Dietmar Offenhuber). Middle: Morais et al. [43]. Right: Liem et al. [38] (Credits: Johannes Liem, Charles Perin, and Jo Wood). For each study, the baseline stimulus is shown on the top and the anthropographic is shown on the bottom.

indicates that an experimental manipulation on a particular design dimension may be expected to promote (resp. reduce) prosocial feelings or behavior, as discussed in [41].

Three main metrics were used in the analysis and interpretation of Boy et al.'s experiment [10]:

- *Self-reported empathy*, measured after each story through 14 questions capturing either empathic concern (with adjectives such as "sympathetic" or "moved") or personal distress (e.g., "disturbed" or "worried");
- *Donation likelihood*, measured after each story by asking how likely the participant was to donate to the cause;
- *Donation allocation*, measured at the end of the experiment by asking to allocate 10 hypothetical dollars to one of the two causes (binary choice).

Contrary to the authors' expectations, no effect of visualization design was found on any of those three metrics. This led them to conduct additional experiments, summarized on Figure 4-top. The figure reports additional information for each experiment, such as the topics used, and the sample size per cell. In EXP 2, the silhouettes from Figure 2-bottom where replaced by generic icons and first names were removed, thereby reducing specificity but making the visualization fully authentic. In EXP 3, a unique anthropomorphic icon was used to represent the entire population, thereby lowering granularity to the baseline level. EXP 4 was similar to EXP 1 but first names were replaced with ages, and less dramatic topics were used. EXP 5 tested a purely textual narrative so it is not included here. Finally, in EXP 6 the text narrative was condensed and in EXP 7, the text became very minimal and age information was removed. Note that the authors were also interested in *expressiveness* as a design strategy, which they defined as the capacity of an icon to convey intentions or emotions [10]. All their anthropographic designs used expressive icons.

No effect of visualization design was found in any of these experiments, though EXP 2 and EXP 7 had some evidence for an effect (resp. negative and positive) on donation allocation. However, the possibility of false positives weakens the evidence.[3] The authors concluded that anthropographic designs are likely not more persuasive than standard visualizations, and while it does not hurt to use them as part of persuasive narratives, they are probably unnecessary.

3.2 Campbell and Offenhuber, 2019

In this study [13], crowdsourcing participants were shown data about the proportion of shelter animals (cats and dogs) who die or go missing. Although the data is about non-human animals, we consider this study as directly relevant to anthropographics. The baseline (design *A*, shown in Figure 3-top-left) consisted of two stacked bar charts – one showing how many animals entered a shelter in 2016, and another showing how many left, broken down into deaths or losses (in orange) and survivals (in green). Other color shades were used to convey the causes of admissions or departures. Three experimental conditions were tested, similar to design *A* but showing different aspects of the data. Design *B* (Figure 3-bottom-left) used a technique called "temporal proximity",

[3] Assuming there is no effect whatsoever, the probability of getting at least one 95% CI that does not cross zero in six experiments is $1 - 0.95^6 = 0.26$. With three uncorrelated dependent variables, the probability rises to $1 - 0.95^{18} = 0.60$, though the three dependent variables were likely correlated.

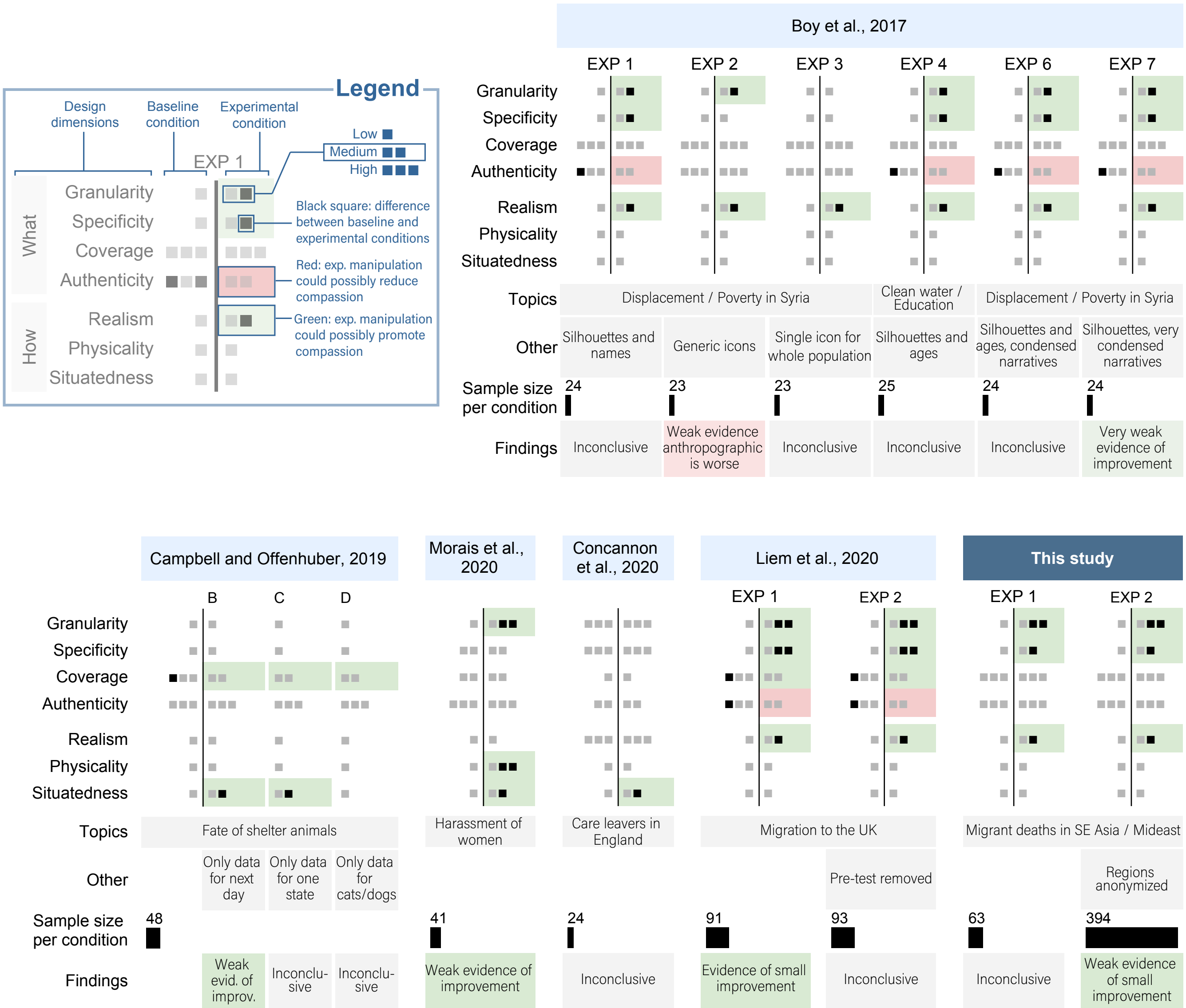

Figure 4: Visual summary of all prior anthropographic studies discussed in our review, and the new study reported in this article.

and presented projected data for the next day. Design *C* used a "spatial proximity" approach and showed only the data from the participant's reported state of residence. Finally, design *D* used a "proximity to interests" approach and showed only the animal species (dogs or cats) the participant reported liking more. Thus, all three experimental designs reduced *coverage* and designs *B* and *C* additionally increased *situatedness* (see Figure 4 for a summary).

The experiment measured:

- *Attitude:* to what extent participants agreed that animal adoption can help decrease the number of animals euthanized in shelters [12, p.59], before and after seeing the visualization.
- *Emotion:* participants were given a standard questionnaire (Geneva emotion wheel) indicating 20 emotions and were invited to report the intensity of the emotions they felt.

The study found evidence that design *B* (temporal proximity) resulted in a change in attitude and an increase in reported disgust, fear and disappointment. All three anthropographic designs were also found to increase interest, which may indicate engagement but not necessarily prosociality. In Figure 4, we thus mark *C* and *D* as inconclusive and mark evidence as weak for design *B* because of the multiplicities involved (e.g., the emotion analysis involved 60 t-tests, 6 of which yielded p<.05).

3.3 Morais et al., 2020

Morais et al. [43] surveyed 28 women who reported having experienced sexual harassment in a public lakeside in Brazil. For each they collected seven pieces of information such as the type of harassment, the time of the day, the perceived age of the perpetrator(s), and the reaction of the victim. The authors then designed two visualizations: a baseline design which presented the data as aggregated statistics (Figure 3-middle-top), and an anthropographic design which turned the data from each woman into a physical object following a plant metaphor (Figure 3-middle-bottom; for an explanation of the encoding, please see [43]). The data physicalization was exhibited in the public lakeside with a legend, while the baseline design was shown as a poster in a nearby park. The characteristics of the two designs are summarized on Figure 4.

In both locations, the authors surveyed people who were exposed to the visualization and people who were not using two metrics:

- a subset of the *self-reported empathy* scale used by Boy et al. [10],
- a *donation allocation* metric, where participants were asked to split a hypothetical monetary reward between themselves and an anti-harassment campaign.

They found weak evidence that the anthropographic design led to higher self-reported empathy than the baseline shown at the remote location, and very weak evidence that it led people to allocate more money. They also found weak evidence that people donate more after seeing harassment plants than those who were at the same location but did not see any visualization. The authors concluded that if an effect exists, it is likely small, and that *"more controlled studies are necessary to understand the effect of anthropographics on compassion"* [43].

3.4 Concannon et al., 2020

Concannon et al. [15] designed and evaluated personalized data-driven video narratives about care leavers in England, i.e., young adults who are transitioning from foster care to an independent life. The video, which showed the story of a fictional care leaver, changed according to where the viewer lives, based on data about local policies for supporting care leavers. Although the video was data-driven, it did not contain any data visualization. We nonetheless include this study because it closely relates to anthropographics. The video was designed to promote empathy by *i)* using a single person to tell a broader story (minimum coverage), and *ii)* selecting data based on where the viewer lives (situated). One of the goals was to find out whether the use of local data (vs. data from a random location) can promote engagement and empathy. The two conditions are summarized in Figure 4. The study failed to find quantitative evidence that situatedness promotes empathy, although qualitative data suggested it may promote engagement.

3.5 Liem et al., 2020

The last study is a crowdsourced study by Liem et al. [38]. In the first experiment, participants were randomly assigned to one of three visualizations of UK migration data. The baseline visualization, shown in Figure 3-top-right, was an interactive flow map that let users explore the extent of in/out migration flows between the UK and any European country, as well as the main stated reasons for migration. The data could also be examined year by year. The anthropographic version (Figure 3-bottom-right) offered users to go through the story of any of six fictional characters, represented as animated icons. The stories included the reasons why the person moved to the UK, what they gained from it, and what they did afterward. Participants then had three minutes to also explore the baseline visualization. Both designs are again summarized in Figure 4. The study also included a third condition using step-by-step explanations. While its intent was to convince people that immigration was not a threat by educating them, it did not use typical anthropographic design strategies so we do not discuss it here.

The main metric of the study was a standard test on *immigration attitudes*, given both before and after exposure to the visualization, and broken down into two submetrics: opposition to immigration (4 questions) and perceived immigration threat (3 questions). There was evidence for a very small within-subject reduction in perceived immigration threat among people exposed to the anthropographic. In order to get more unbiased responses, the authors replicated the experiment without the pre-test only using between-subjects differences on the post-test as a measure of effect size. No difference was found between the anthropographic and the baseline. Interestingly, the third (step-by-step) condition yielded a large between-subjects increase in perceived threat.

The authors concluded that the study failed to provide evidence that their two designs can improve attitudes towards immigration, and while this does not prove that no other design could work, researchers need to *"gather more empirical evidence before making strong claims regarding the benefits of storytelling in visualization"*.

4 EXPERIMENT 1: COMBINED DESIGN STRATEGIES

Experiment 1 aims at measuring the effect of anthropographics in a situation where this effect is likely to show up: when multiple anthropographic design strategies are combined. Given previous failures to detect an effect (see section 3), we decided to start with an experiment that tries to maximize effect size and the likelihood of finding an effect at the expense of explanatory power. Both our experiments have been approved by the comité d'éthique de la recherche Université Paris-Saclay, reference CER-Paris-Saclay-2019-006.

4.1 Anthropographic Design Rationale

We contrast the five studies we reviewed in the previous section with our own. However, at the time we ran our experiments (Oct and Nov 2019), only Boy et al. [10] had been published. Our study is thus based on theirs. Like Boy et al., our study manipulates *granularity*, *specificity*, and *realism*, all of which can be hypothesized to promote prosociality [42]. However, we thought the negative results in Boy et al. were likely due to the limited granularity and limited amount of real personal information conveyed in their designs. Therefore, we chose to maximize *granularity* as well as *authenticity*. To summarize, our anthropographic designs have:

- *Maximum granularity.* This strategy uses marks that represent one individual each, connecting the reader to the data (and fate) of specific persons instead of an aggregate of people. As Figure 4 summarizes, Boy et al. only used medium granularity in five of their experiments, with marks that each represent 1% of the population visualized, and low granularity or text in the remaining experiments. Campbell & Offenhuber used low granularity (bar charts) whereas Liem et al. and Morais et al. used both maximum granularity by telling the story of some selected individuals.

- *Intermediate realism.* This strategy, which was also used by Boy et al. and Liem et al., employs human icons to reinforce the idea that the data is about people. In contrast with Boy et al., however, each icon in our case maps to a single person, in a way that is fully consistent with a maximum granularity approach.
- *Medium specificity.* This strategy uses distinctive marks which communicate attributes of groups or individuals. It has been hypothesized that more distinctive marks may increase the likelihood of prosocial behavior [42]. There is however a constraint, because highly specific marks may result in the persons whose data are visualized being identifiable which would violate their privacy. For this reason, Boy et al. and Morais et al. used medium specificity, Campbell & Offenhuber used low specificity, and only Liem et al. used high specificity and addressed the privacy issue using partial authenticity.
- *Full authenticity.* This strategy uses marks which do not add any synthetic data to the visualization. As discussed in section 3, partial authenticity was common in many of the previous experiments (see Figure 4), for example, by adding name and age labels (Boy et al. and Liem et al.), using a unique silhouette for each mark (Boy et al.) or showing a fictional person/actor (Concannon et al.). We believe that using partial authenticity carries the risk of reducing prosocial dispositions and we therefore decided to only use full authenticity.

In an attempt to maximize effect size, we combine the four design strategies. This means that we use *maximum granularity*, such that individual people appear, *intermediate realism*, to reinforce the idea that each data point is a person, *medium specificity*, such that the reader can learn something pertinent about each person without rendering that person identifiable, and *full authenticity*, such that the reader cannot dismiss the information presented as being fake and potentially exaggerated. We henceforth use the term **information-rich** as a shorthand to refer to anthropographic designs that have such characteristics (including also high specificity). Meanwhile, designs that are low on one or several of these dimensions, such as classical statistical charts, will be referred to as **information-poor**.

Before detailing the designs we use, we note that other design strategies can be expected to promote prosociality. Past studies have for example manipulated situatedness with this intention (e.g., Campbell & Offenhuber, Morais et al.). We chose to focus on manipulations similar to Boy's study but adapted in order to increase effect size, and that can be tested on an online crowdsourcing platform.

4.2 Stories and Datasets

In this experiment, participants were shown information about migrants deaths. We chose this topic for two reasons. First, designers typically use anthropographics to convey data about large-scale human suffering, such as caused by human rights violations or natural catastrophes. Our topic is thus similar to stories used in journalistic articles and charitable giving projects.

Second, there exist detailed data on this topic that allows us to visualize authentic information about individuals. The data we use comes from the *Missing Migrants Project*[4], which *"tracks deaths of migrants, including refugees and asylum-seekers, who have died or gone missing in the process of migration towards an international destination"*. Each dataset row describes an incident, such as drownings, vehicle accidents, or shootings. Every incident contains the number of victims and survivors (split into men, women, children, or unknown), where and when the event occurred. The data is collected from official records and media reports[5].

We extracted two subsets from the Missing Migrants dataset: one focusing on migrants in Southeast Asia in 2018, and one focusing on migrants in the Middle East in 2018. These are two regions with a similar number of fatalities, which allowed us to create two stories roughly comparable in gravity: ***story 1*** about migrants in Southeast Asia, and ***story 2*** about migrants in the Middle East. The accompanying textual narratives were minimal, as in the last two experiments from Boy et al. [10] and following their recommendation. They will be shown in the next section.

4.3 Visualization Designs and Stimuli

The experiment involved two visualization designs: an information-rich anthropographic design (which we refer to as ***infoRich***), and an information-poor design acting as a comparison baseline (***infoPoor***). Both designs were used to visualize both stories, leading to four different stimuli: ***(story 1, infoPoor)***, ***(story 1, infoRich)***, ***(story 2, infoPoor)***, and ***(story 2, infoRich)***. Figure 1 shows the stimuli *(story 2, infoPoor)* and *(story 2, infoRich)*. All four stimuli are available in the online supplemental material.

The *infoPoor* baseline visualization (Figure 1-left) is a simple bar chart showing the number of migrants who died or survived an incident during their migration attempt in 2018. We use a bar chart as the baseline condition because it is a common design choice to represent statistical data. It is also a widely recognized visualization type even among laypeople. This is a simple visualization with *low granularity* (since it shows aggregated data), *low specificity* (since no information is shown about people except for their survival status), and *low realism* (since the two bars do not look like humans).

The *infoRich* visualization (Figure 1-right and Figure 5) shows one row for each incident that took place in 2018. A text label indicates the date, nature, and location of the incident. Underneath, each individual migrant involved in the incident is shown as a human silhouette conveying their gender according to the dataset (male, female, unknown), age group (child or adult), and whether they died or survived (red vs. light pink). Participants had to scroll to see the entire visualization and press the "Next" button. The inset on the right of Figure 1 (not shown in the experiment) gives an indication of the full height of the visualization. This visualization has *maximum granularity* (since each mark is a person), *intermediate specificity* (since some information is shown about each person but

[4] https://missingmigrants.iom.int
[5] See missingmigrants.iom.int/methodology for more on the methodology.

Figure 5: An excerpt of the information-rich visualization showing data from the Middle East from February 13 to April 21, 2018. For the legend see Figure 1.

not enough to make them unique or identifiable), and *intermediate realism* (since iconic anthropomorphic representations are used).

Being fully authentic, the *infoRich* design has the same level of authenticity as the *infoPoor* design, and thus the experiment does not manipulate this design dimension. Even though the marks are not drawings of real people, we consider them fully authentic because all attributes that vary across marks and thus encode information (e.g., age and gender) are based on real data. Section 4.5.1 in Morais et al. [42] further discusses how visualizations with intermediate realism can be considered fully authentic. Also note that as can be seen in Figure 5, for some events, only the number of victims and survivors is known whereas for others, we have additional data on age group or sex. While missing data reduces specificity, the *infoRich* design is still much higher in specificity than the *infoPoor* design.

The three major design dimensions on which the *infoPoor* and *infoRich* designs differ are summarized visually at the bottom right of Figure 4. The two visualization designs differ in other respects – importantly, *infoRich* requires a significant amount of scrolling, while *infoPoor* fits in the browser window. Similarly, *infoRich* requires considerably more time to be fully examined than *infoPoor*. All these potential confounds will make it hard to identify the cause of an experimental effect should we find one. However, as we already mentioned, the simultaneous manipulation of multiple design characteristics is a deliberate feature of this experiment, whose goal is to confirm that we are indeed able to detect an effect if we try to amplify it as much as possible.

4.4 Experiment Design

Each participant saw *story 1* (Southeast Asia migrants) followed by *story 2* (Middle East migrants), always in that order. However, participants were randomly split in two groups (see also Figure 6): the ***poorFirst*** group saw the *(story 1, infoPoor)* stimulus followed by *(story 2, infoRich)*, while participants in the ***richFirst*** group saw *(story 1, infoRich)* followed by *(story 2, infoPoor)*. In other words, all participants saw the two same stories in the same order, but some participants saw the *infoRich* design followed by the *infoPoor* design, while others saw the two designs in reverse order.

Thus, the experiment followed a mixed design. The story (*story 1* or *story 2*) and the visualization design (*infoPoor* or *infoRich*) were within-subject independent variables, while the order of presentation (*poorFirst* or *richFirst*) was a between-subjects independent variable.

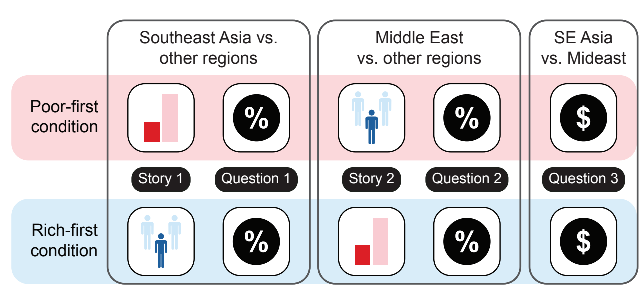

Figure 6: Visual summary of the design and procedure.

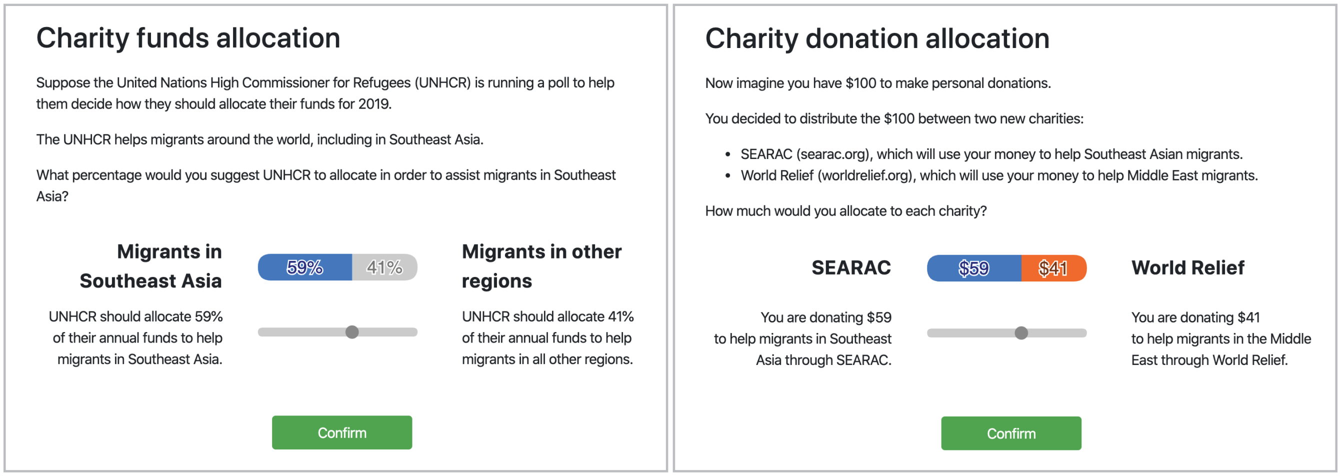

Figure 7: Two pages from experiment 1 asking participants to allocate hypothetical money: on the left, *question 1*, and on the right, *question 3* (see text for details).

4.5 Measurements

We captured prosociality indirectly through questions involving allocating hypothetical money between two options. Each participant had to answer three monetary allocation questions:

- ***Question 1*** (Figure 7-left) asked participants to suggest how much of its funds should a charity allocate to migrants from *story 1* (Southeast Asia) compared to migrants in other regions of the World. The charity was a real organization, briefly described before the question (more details in section 4.6).
- ***Question 2*** asked participants to allocate funds from another charity between migrants from *story 2* (Middle East) and migrants from other regions of the World.
- ***Question 3*** (Figure 7-right) asked participants to allocate a fixed amount of hypothetical personal money between a charity that helps migrants from *story 1* (Southeast Asia) and a charity that helps migrants from *story 2* (Middle East).

We chose to use multiple questions in order to test different measurement strategies and inform our future experiments. Importantly, instead of asking participants how much money they were willing to give to a single charity, all three questions asked them to allocate money such that the same fixed amount will be spent helping migrants regardless of their response. Such a question framing is likely less prone to social desirability bias [23] and ceiling effects (i.e., there is no cost involved in donating arbitrarily high amounts of hypothetical money). In an attempt to further reduce social desirability bias, *question 1* and *question 2* asked participants to make decisions over other people's money rather than their own. The way the two questions were framed also discouraged participants from allocating the same amount of money to the two options (as might happen in *question 3*), since the two options partition the world unequally (Southeast Asia or the Middle East vs. other regions). It is important to note that the three questions do not measure *prosociality* per se (all the money is used to help migrants regardless of the response), but instead participants' *judgment* about the extent to which migrants in *story 1* or *story 2* need help. Our reasoning is that if, in real life, an anthropographic is able to convince people (e.g., in a fund raising situation), that the persons represented need help, it will also likely promote prosocial behavior.

From those three questions, we define three ways of measuring how a participant allocates money:

- The ***relative allocation*** $\in$ [-100, 100] is the difference between a participant's response to *question 1* and their response to *question 2* (i.e., *question 1* − *question 2*).[6] This measure captures participants' preference for helping migrants from *story 1* compared to those in *story 2*, and serves to define our primary outcome.
- The ***single-story allocation*** $\in$ [0, 100] is the participant's response to *question 1*.
- The ***comparative allocation*** $\in$ [0, 100] is the participant's response to *question 3*.

Finally, based on these three measures, we define different ways of capturing average participant bias towards the information-rich visualization design:

- The ***information richness effect on relative allocation*** $\in$ [-200, 200] is the mean difference in *relative allocation* between the *richFirst* and the *poorFirst* groups (*richFirst* − *poorFirst*). If more resources are allocated to the story presented with the information-rich visualization irrespective of it being shown first or second, this value should be strictly positive. Alternatively, if information richness has no effect, the value should be zero. This measure is our primary outcome because we expected that a within-subject measure would remove some variability across participants and lead to more precise estimates [17, Chap. 8].
- The ***information richness effect on single-story allocation*** $\in$ [-100, 100] is the mean difference in *single-story allocation* between the *richFirst* and the *poorFirst* groups. Although

[6]For all questions, we consider the response to be the value selected for the left option. For example in Figure 7, the responses are 59% (left image) and $59 (right image).

the previous effect size has the potential benefit of providing more precise estimates, it can potentially yield biased answers since participants were exposed to both stories and the both types of visualizations and are thus potentially aware of the experimental manipulation [44]. This is not the case for this second metric, since at the time participants answer *question 1* they have only seen *story 1* and a single visualization design (more details on the sequencing in section 4.6). Thus, this between-subjects measure may be less sensitive but has more internal validity. We therefore include it as a secondary outcome.

- The ***information richness effect on comparative allocation*** $\in$ [-100, 100] is the mean difference in *comparative allocation* between the *richFirst* and the *poorFirst* groups. We use this metric as another secondary outcome. Our goal was to test a question involving donation rather than fund allocation, and which forces a direct comparison between *story 1* and *story 2*.

These three effect size metrics provide three different ways of answering our research question *"does an information-rich anthropomorphic visualization of humanitarian data increase donation allocations compared to a simple bar chart?"*. They were formulated and prioritized before we conducted the experiment.[7]

4.6 Procedure

Participants were asked to navigate through a series of pages.[8] They were not allowed to revisit previous pages. The first five pages had general instructions, a consent form, and a fund allocation example similar to the one shown in Figure 7-left, whose role was to familiarize participants with the type of question they will see and the user interface.

The actual experiment started on page six and is summarized in Figure 6. Page six showed a short textual narrative with a visualization (either *(story 1, infoPoor)* or *(story 1, infoRich)* depending on the condition). The two stimuli were similar to those on Figure 1 except they featured Southeast Asia. The next page was a brief textual presentation of the United Nations High Commissioner For Refugees (UNHCR), informed by their official website. The next page, shown in Figure 7-left, asked participants to select a value between 0% and 100% to suggest how should UNHCR split its annual funds to help migrants in Southeast Asia vs. other regions in the World (*question 1*). On the following page, the participant was asked to briefly justify their answer in a text field.

The next four pages (pages ten to thirteen) were similar, except they corresponded to *story 2*. The featured region was the Middle East (see Figure 1), and the featured organization was the International Organization for Migration (IOM). After seeing both visualizations, *question 3* asked participants to allocate $100 between a charity helping migrants in Southeast Asia and another one focusing on migrants in the Middle East. The last pages from the experiment had an attention check question and a general information form.

[7] The preregistration of experiment 1 can be accessed at https://osf.io/epzub.

[8] Screenshots from both experiments can be found in our accompanying material in the OSF repository. Additionally, both experiments can be tested anonymously and untracked on this github page: anthropographics.github.io.

4.7 Experimental Platform

We ran the experiment on the Prolific[9] research crowdsourcing platform. The job was titled "Study on judgment and humanitarian issues", and it asked participants to read facts about humanitarian issues and allocate hypothetical money. Participants who accepted the job were redirected to an external page hosted by the authors.

Only desktop and laptop computer users were allowed to participate, and the size of all experiment pages was fixed to 900×600 pixels to ensure that all participants had to scroll through the information-rich visualization in the same fashion. If the browser window was smaller than this size or used a zoom factor less than 100%, participants were instructed to adjust their window size or zoom factor.

The experiment was only available to contributors who did not participate in our pilots, were fluent in English, and had at least a 95% job approval rate. Contributors who failed the attention check or reloaded the page were also excluded from the experiment (they were warned in advance not to do so).

Contributors were offered a reward of 1.20£ for an estimated (and actual) completion time of eight minutes. The time estimation was based on multiple in-person and Prolific pilots, which also served to ensure the clarity of our instructions.

4.8 Participants

Our planned sample size was $N = 128$ (64 per condition). We chose this sample size to achieve a power of 0.8 to detect a "medium" Cohen *d*'s effect size of 0.5, as computed by the G*Power software [22].

We got valid data from $N = 128$ participants ($N = 64$ for *infoPoor* and $N = 64$ for *infoRich*). Participants were 33% female, with a mean age of 30. They were from 20 different countries, mostly from Europe (top-3 were UK 33%, Portugal 15%, and Poland 10%).

4.9 Planned Analysis

All analyses presented in this section were pre-registered (OSF link in section 4.5). We report all our results using estimation statistics [16, 20] and draw our inferences from graphically-reported point estimates and interval estimates [18, 35]. All effects are estimated using BCa bootstrap confidence intervals, which provide reasonably accurate interval estimates without distributional assumptions for sample sizes of 20 or more [33] (the current experiment has 64 participants in each condition). Since we identify a single primary outcome in this analysis, we do not adjust for multiplicity [5, 37].

4.9.1 Attrition. Out of the 144 contributors who agreed to the consent form, 16 did not finish the experiment, which corresponds to an attrition rate of 11%. Among those, 50% failed the attention check, 12% reloaded the page, and 38% quit for unknown reasons. We compared attrition rates between the two experimental groups to make sure participants did not drop out considerably more often in one condition than the other, which would constitute a threat to validity [58]. The attrition rate was 14% in the *poorFirst* group and 9% in the *richFirst* group, and the difference was -5%, 95% CI [-16%, 6%]. Thus we found no evidence that attrition rate differed substantially across experimental groups.

[9] www.prolific.co

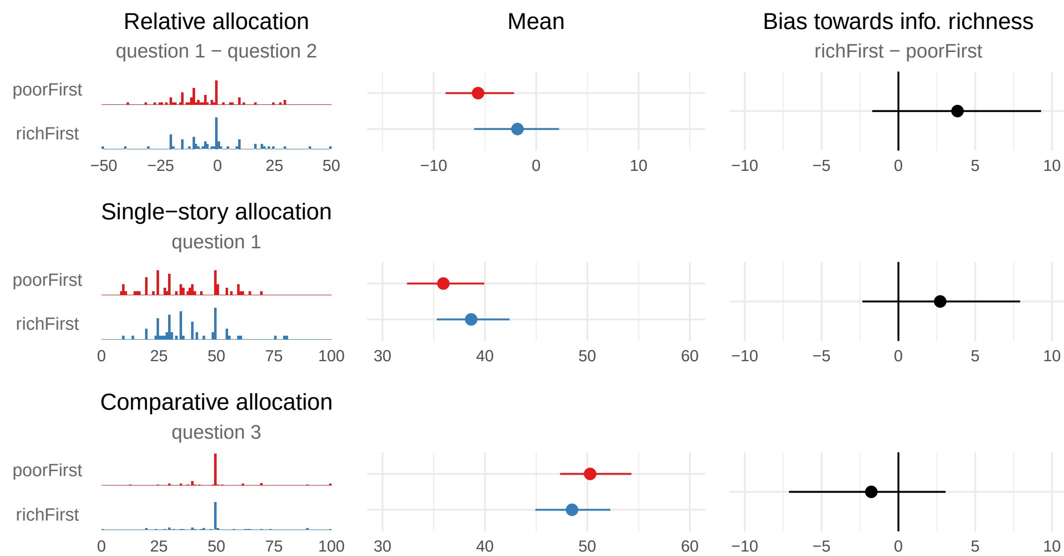

Figure 8: Raw responses, mean responses, and effect size for the three effect size metrics in experiment 1, with 95% CIs.

4.9.2 Information richness effect on relative allocation. This is our primary way of capturing the average participant bias towards the information-rich design, and it is defined as the mean difference in *relative allocation* (i.e., answer to *question 1* – answer to *question 2*) between the *richFirst* and the *poorFirst* groups (see section 4.5 for details). Results are reported on the top row of Figure 8, with the left chart showing the distribution of *relative allocation* responses, the middle chart showing *richFirst* and *poorFirst* means, and the rightmost chart showing the effect size, the *information richness effect on relative allocation.* Although the effect size interval mostly covers positive values, the evidence is insufficient to conclude that there is an effect.

4.9.3 Information richness effect on single-story allocation. This alternative way of capturing bias towards information richness is the mean difference in *single-story allocation* (i.e., answer to *question 1*) between the *richFirst* and the *poorFirst* groups (see section 4.5). Results are reported in the second row of Figure 8. Again, there is insufficient evidence to conclude that that there is an effect. Note that in section 4.5 we expected the previous metric to yield a more precise estimate than this one, but this did not happen.

4.9.4 Information richness effect on comparative allocation. Finally, this third measure of bias is the mean difference in *comparative allocation* (i.e., answer to *question 3*) between the *richFirst* and the *poorFirst* groups. Although this time the interval admits more negative values, there is still no evidence to suggest that the effect is either positive or negative.

In summary, the data collected in this first experiment does not support the hypothesis that our information-rich anthropographic design increases donation allocation compared to a simple bar chart (neither does it support the opposite hypothesis). The data suggests, however, that should there be an effect, it is likely smaller than a difference of 10% (see CIs in Figure 8).

4.10 Additional Analyses

In addition to the planned analyses presented so far, we examined the distributions of responses to the three monetary allocation questions (see histograms for *question 1* and *question 3* in Figure 8). We saw that participants were inclined to allocate donations evenly. For *question 3*, a remarkably large number of participants split their funds evenly between *story 1* and *story 2*. Even in *question 1* and *question 2* where participants had to split funds between a single geographical region and the rest of the world, most participants allocated between 30% and 50% to the single region.

Analyzing the open-ended justifications for allocation decisions helped us further understand what happened: most participants who split their money equally in *question 3* invoked fairness as a reason. For example, participants wrote: *"I want to equally help both as much as I can. It may come off strangely if we think of the circumstances and data given, but both are in need and the thought of putting one above the other is bad for me"*; *"I don't think that migrants from one country or the other deserve more my donation."*; *"Fairness, I know that people in both regions are suffering and wanted to provide support to all people."*.

While examining participants' justifications for their responses, we also found that many of them were influenced by their prior attitudes. Sometimes those attitudes arose from factual knowledge, e.g., *"The crisis in the middle east is major. I believe 30% of all available budget should be directed there to help the people in need. But there are other part of the world that needs help, then can use the other 70%"*. Often, though, attitudes were influenced by subjective feelings or prior experience. For example *"I lived in Thailand for four years. The migrant crisis in that part of the world is important to me as I witness it a few times"*; *"I also have a desire that I do want to help those in Southeast Asia, as I have spent time there"*; *"I feel more sympathy towards Asian migrants"*.

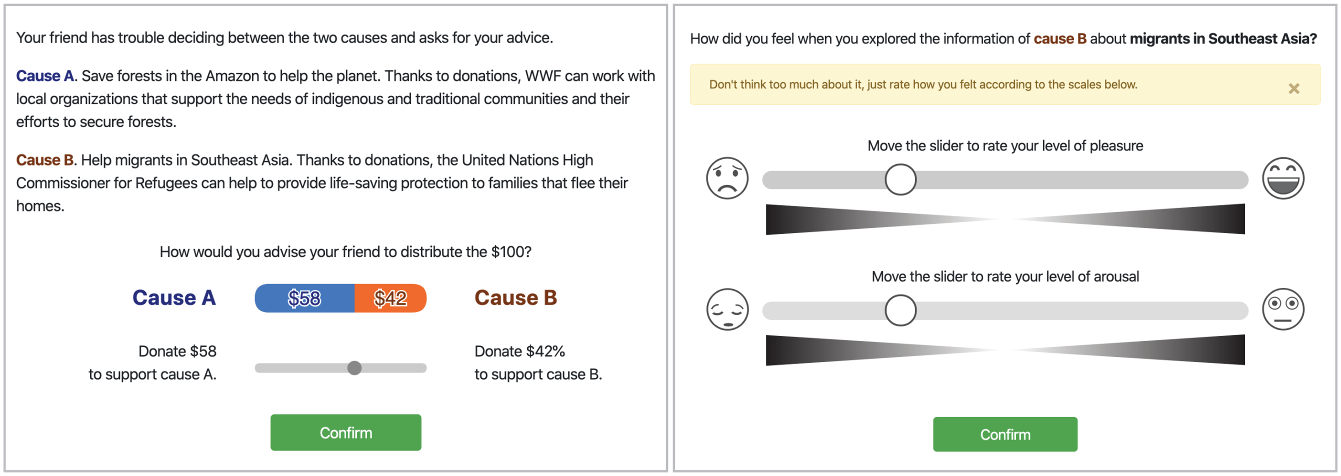

Figure 9: Two pages from experiment 2: the donation allocation question (left) and the affect questions (right).

4.11 Discussion

We did not find evidence to support the effectiveness of the information-rich visualization design, both according to our primary measure, and our two secondary measures. There can be several reasons for this. One of course is that the anthropographic design is ineffective at promoting prosociality, another is that its effect (be it positive or negative) is smaller than our experiment was able to measure. In this case, a larger sample size would be needed to detect it.

Our examination of the distribution of responses showed that participants tended to allocate funds or donations evenly. This tendency, which has been observed in a previous anthropographics study [10], likely reflects a naive diversification bias [4] whereby people prefer to allocate limited resources evenly across options.[10]

Participants' prior attitudes might also have contributed in overshadowing a possible effect of the anthropographic. As revealed by the free-text justifications, many participants were influenced by prior knowledge, prior experience or prior feelings towards a group. Therefore, not naming regions could help homogenize responses and thus make the effect of anthropographics stand out.

5 EXPERIMENT 2: INCREASING STATISTICAL POWER

Experiment 2 was designed to overcome the limitations we identified in experiment 1. It differs in six major ways:

(1) We dropped story as an independent variable. This is because in section 4.9, we found no evidence that the within-subject effect size (*information richness effect on relative allocation*) yields more statistical power than the between-subjects one (*information richness effect on single-story allocation*). As a result, we only kept the Southeast Asia story and dataset.
(2) We use a question requiring money allocation between two options that are difficult to value equally. We expect this to reduce the naive diversification bias.
(3) We frame the donation question differently, by asking participants to advise a friend. This is done to reduce social desirability bias and to allow participants to answer independently from their socio-economic status (one participant in experiment 1 commented that they would not be able to donate money).
(4) We include a condition where the geographical regions are unnamed, with the expectation that this will reduce the influence of prior attitudes.
(5) We additionally measure affect (arousal and valence), in order to investigate whether information-richness can elicit emotional responses irrespective of its effect on decision making.
(6) Finally, we use a much larger sample size in order to maximize our chance of observing an effect.

5.1 Stories and Datasets

The three monetary allocation questions from experiment 1 asked participants to split money between two similar causes, both involving migrants in a region of the World. This seems to have encouraged participants to allocate money evenly, especially in *question 3*. In order to find two options that make it harder to justify a 50/50 donation, we conducted a pre-study where we asked participants to allocate money between the Southeast Asia story (without the visualization) and one of the four following stories (all provided in textual form only):

- *Supporting earthquake victims.* The story asked for help to rebuild the lives of people who experienced an earthquake that took almost 9,000 lives.
- *Helping to prevent Zika virus outbreaks.* The story called participants to help prevent a new Zika virus outbreak to avoid damages to pregnancies.

[10] Equal allocation is naive when it ignores information about the different options. For example, it is generally not fair to allocate the same funds to two regions of very different population size.

- *Helping to clean up the oceans*. The story asked for help to clean up the oceans to avoid the death of marine animals.
- *Saving forests*. The story called for saving forests that are the habitat of indigenous people and are consumed by fires.

All four stories were inspired by real news events, but region names for these four stories and the migrants story where anonymized to reduce the influence of prior attitudes. We ran the pre-study on the Prolific platform. A total of 154 participants were randomly assigned to one of the four stories, and asked to split hypothetical money between the tested story and the migrant story. Upon examining the distributions of responses, we selected the "saving forests" story because it yielded the smallest spike in 50/50 donations (13% of all responses, compared to 18% for cleaning up the oceans, 22% for fighting Zika virus, and 27% for helping earthquake victims).

Throughout this section, we refer to the two stories according to the order in which they appear in the experiment, i.e., as ***story 1*** for saving forests and as ***story 2*** for migrants in Southeast Asia. Please note that these names refer to different stories than in experiment 1.

5.2 Visualization Designs and Stimuli

This experiment involved the same information-poor and almost the same information-rich visualization designs. Accordingly, two different visualizations were created to convey the data of *story 2* (see again Figure 1 and Figure 5 for similar visualizations, but of a different dataset)[11]. For each of these two visualizations, a variant was created that was the same except all information allowing to identify the Southeast Asia region was removed. This led to a total of four different stimuli: ***(infoPoor, named)***, ***(infoPoor, anonymized)***, ***(infoRich, named)***, and ***(infoRich, anonymized)***.

5.3 Experiment Design

The experiment followed a 2×2 full factorial design with visualization design (*infoPoor* or *infoRich*) and anonymization (*named* or *anonymized*) as the two independent variables. Both were between-subjects, so each participant only saw one of the four stimuli.

5.4 Measurements

Our primary research question was again: "*to what extent does an information-rich visualization design affect donation allocations compared to an information-poor design?*". We used this time a single question asking participants to split a hypothetical donation of $100 between the two causes mentioned previously. Thus we have a single measure of effect size which we refer to as the ***information richness effect on comparative allocation*** $\in$ [-100, 100], and which is the mean difference in the money allocated to *story 2* between all participants in the *infoRich* condition and all participants in the *infoPoor* condition. In other words, this measure is a contrast that collapses four conditions into two, looking at the main effect of visualization design and ignoring anonymization.

We also had two secondary research questions:

(1) *Does the effect depend on whether regions are anonymized?*
(2) *To what extent does an information-rich design have an influence on reported affect compared to an information-poor design?* We measured affect using the Affective Slider [6], a continuous and modern version of the Self-Assessment Manikin scale [11]. This scale is subdivided into ***valence*** $\in$ [0, 1], which captures how positive or negative the stimulus is perceived to be, and ***arousal*** $\in$ [0, 1], which captures how calming or exciting it is.

In addition, we framed two auxiliary research questions detailed in section 5.7.5. As before, all research questions and measurements were decided before we ran the experiment.[12]

5.5 Procedure and Platform

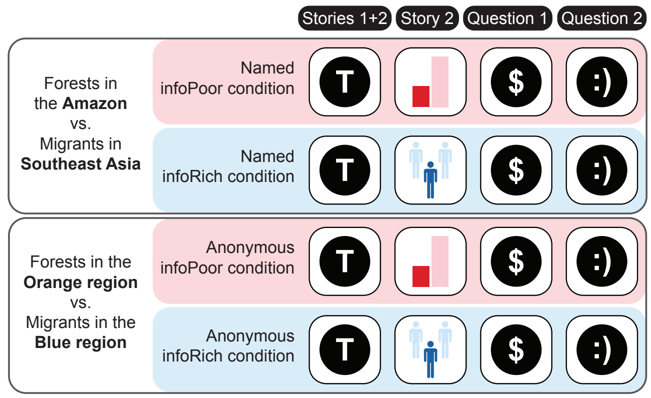

Figure 10: Visual summary of the design and procedure of experiment 2.

Compared to experiment 1, the experiment was simplified and shortened—participants only had to read initial instructions and accept the consent form before they proceeded with the task. The experimental task spanned four pages, summarized in Figure 10. The first page showed both *story 1* (top half of the page) and the textual portion of *story 2* (bottom half). In the anonymized version, the Amazon in *story 1* was replaced with the Blue Region and Southeast Asia in *story 2* was replaced with the Orange region. The second page presented the visualization portion of *story 2* (i.e., one of the four stimuli described in section 5.2). The third page contained the donation allocation question (Figure 9-left). The question was framed similarly as *question 3* from experiment 1, except participants were asked to imagine advising a friend instead of making the donation themselves. The fourth page contained the affect scale (Figure 9-right) which was based on the code and guidelines proposed by the Affective Slider authors[13].

Experiment 2 also ran on the Prolific platform using the same infrastructure as experiment 1. We used the same quality control strategies described in section 4.7 and screened out participants from the previous experiment. Contributors were offered a reward of 0.45£ for an estimated (and actual median) completion time of three minutes.

[11] Neither of the two information-rich visualizations indicates the exact location of the incident as was done in experiment 1 (as visible in Figure 5) to ensure the same level of information richness between the anonymous and non-anonymous version of the information-rich visualization. The complete experiment pages including the respective visualizations can be accessed at anthropographics.github.io/.

[12] The pre-registration of experiment 2 can be accessed at https://osf.io/c82hn.

[13] Affective Slider Github page: https://github.com/albertobeta/AffectiveSlider.

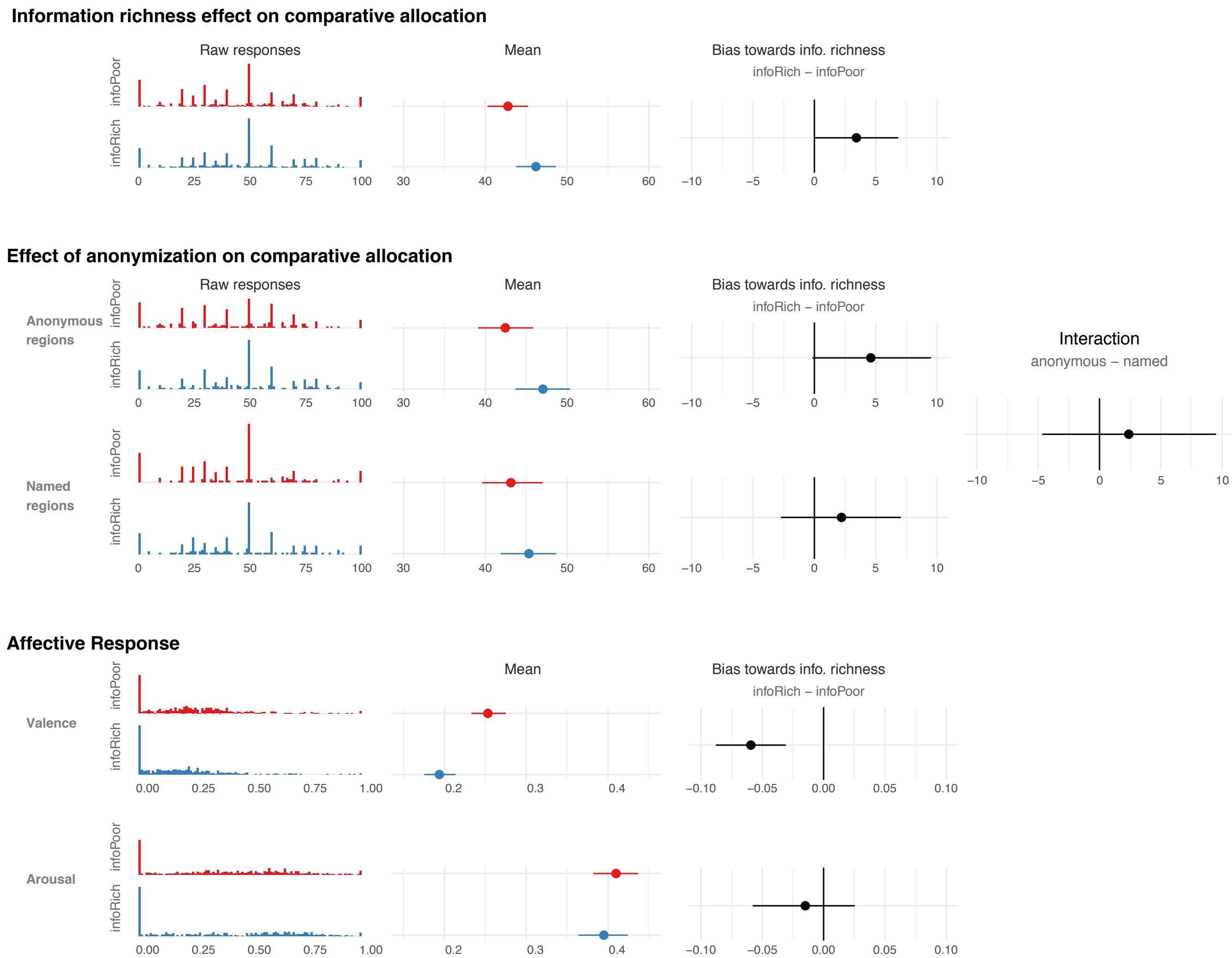

Figure 11: Raw responses, mean responses, and effect size in experiment 2, with 95% CIs.

5.6 Participants

Our planned sample size was $N = 788$ ($N = 197$ per condition). This sample size was chosen to provide a statistical power of 0.8 to detect a "small" Cohen d's effect size of 0.2 for our first research question, where the four conditions are collapsed into two. We received valid data from $N = 786$ participants: $N = 188$ for the *(infoPoor, named)* condition, $N = 207$ for *(infoPoor, anonymized)*, $N = 195$ for *(infoRich, named)*, and $N = 196$ for *(infoRich, anonymized)*. Participants were 49% female, with a mean age of 33. They were mostly from Europe and the United States.

5.7 Planned Analysis

Similar to the first experiment, we report and interpret all our data using interval estimation, and use BCa bootstrap confidence intervals. All analyses were planned before data were collected.

5.7.1 Attrition. Out of the 810 participants who agreed to the consent form, 24 did not finish the experiment, which is an attrition rate of 3%. Among those, 54% failed the attention check, 12% reloaded the page, and 33% quit for unknown reasons. We found no evidence for a difference in attrition rates between conditions, which were 4% in *(infoPoor, named)*, 3% in *(infoPoor, anonymized)*, 3% in *(infoRich, named)*, and 1% in *(infoRich, anonymized)*.

5.7.2 Information richness effect on comparative allocation. This primary outcome answers to what extent an information-rich anthropographic affects donation allocation compared to an information-poor design. The results are reported at the top of Figure 11. There is some evidence of an effect in favor of the information-rich design. This design seems to prompt more donations on average, though the difference is likely less than 7%.

5.7.3 Effect of anonymization. As secondary research questions, we estimated the same effect as before but for the *named* and *anonymized* conditions separately. The results are reported in the middle of Figure 11. There is some evidence that the information-rich design increases donations when the stories feature anonymized regions, while there is no evidence of such an improvement when the regions are named. However, judging by the interaction effect (reported on the right of Figure 11), we cannot conclude that the effect of information richness is affected by anonymization [24].

5.7.4 Affective response. Results for valence and arousal are shown at the bottom of Figure 11. Both scales are between 0 and 1 and centered at 0.5 (meaning that a valence less than 0.5 corresponds to a negative emotion). It can be seen that zero was a common answer for both questions, perhaps because a different response could be interpreted as admitting some level of pleasure or arousal (see again the question framing in Figure 9-right). We have very strong evidence for a difference in reported valence between the two visualization designs: the information-rich design prompted participants to report more negative feelings on average. However, the effect on reported arousal (i.e., the reported intensity of the feelings) is inconclusive.

5.7.5 Auxiliary questions. Our two auxiliary research questions were: *"how strong is the correlation between donation and affect?"* and *"does anonymization have an overall effect on donation behavior?"*. The correlation between valence and donation was negative but small ($\tau \approx -0.1$), and we found no evidence for a correlation between arousal and donation, nor for a main effect of anonymization on donation. The full analyses can be found in the supplementary material.

6 GENERAL DISCUSSION

Our first experiment was designed to detect a medium effect (d = 0.5) of information richness on prosociality. Inconclusive results led us to run a second experiment with an increased sample size to detect an effect as small as d = 0.2. We also revised measurement questions to reduce the impact of naive diversification strategies and prior attitudes. This second experiment allowed us to find some evidence of an effect, albeit small.

A small effect can be relevant in practice, as designing a communicative visualization is likely inexpensive in comparison to other costs typically involved in large donation campaigns, and even a modest increase in individual donations can have concrete consequences for individual people. At the same time, the effect we measured is substantially smaller than we (and perhaps many designers would have) expected, which poses serious difficulties for research. Indeed, our second experiment was barely able to provide sufficient evidence, despite having involved about 800 participants for a cost of about 500 GBP (550€ or $650 in mid-2020). As this effect arises from three design strategies combined and other potentially influential factors (e.g., asking participants to scroll, different exposure times), disentangling these effects may require sample sizes that are impractical to achieve.[14]

We found clear evidence that visualization design affected participants' reported affect, with participants reporting more negative emotions when exposed to the information-rich anthropographic. Although this was not our primary research question, this finding suggests that our anthropographic design may have a moderate but tangible psychological effect on people, irrespective of their monetary allocation decisions. This is in line with designers' intuitions about the potential of anthropographics for evoking empathy [28, 36, 57], even if empathy is not necessarily a good predictor of actual helping behavior (as discussed in section 2.1). Nevertheless, it could be interesting for future experiments to disentangle the causes of the observed change in reported affect, since only a subset of our design strategies may be responsible for it.

There are several important limitations to our study. First, as in almost any study, our measures are not the constructs they are designed to capture. Our monetary allocation questions are proxies for prosociality based on participants' judgments about the extent to which the people visualized need help from charities. There is likely a strong connection between the two, but we cannot exclude the possibility that participant responses were tainted by social desirability biases [23] or demand characteristics [44]. Regardless of the efficacy of anthropographics, the validity of future studies would benefit from the observation of actual participant behavior, either elicited explicitly or observed in the wild in more realistic situations. An example of the the former case is asking participants to donate back part of their compensation for participation as done by Small and colleagues [52], while for the latter Concannon and colleagues [15] have suggested embedding studies in real charity campaigns.

Another key limitation is that we tested a single anthropographic design. It is of course possible that this precise design is ineffective, while there are other designs that have an effect on participants. The design space we have used features other relevant dimensions such as coverage and situatedness, which could be interesting to explore further (as was done by Campbell and Offenhuber [13]). However, it is unlikely that previous studies or the design space by Morais et al. [42] cover all possibilities. We nevertheless hope that our review of past experiments and the diagrams we developed in section 3 will help researchers design new studies.

7 CONCLUSION

There is increasing interest in anthropographics, i.e., visualizations designed to make readers compassionate with the persons whose data is represented. Empirical studies have started to test designers' recommendations and intuitions by examining whether various design strategies indeed promote prosocial feelings or behavior. Results have been mostly disappointing so far, in contradiction with the intuitions of many designers and researchers.

This work contributes a detailed overview of past experiments and introduces two new experiments that use large samples and a combination of design strategies to maximize the possibility of finding an effect. Our anthropographic design had a clear but modest effect on reported affect, and a small effect at best on money allocation. Such a small effect may be relevant for large-scale donation campaigns, but it seems very difficult to study experimentally.

It remains possible that there exist alternative anthropographic design strategies that have a large effect on measures of affect, decision making, or behavior. However, the overall inconclusive results from the range of experiments conducted by researchers so far call for some skepticism. There is presently no clear evidence that if designers employ current anthropographic design strategies,

[14] The relationship between sample size and effect size is not linear. For example, being able to detect an effect four times smaller than experiment 2 (d=0.05) while maintaining a power of 0.8 would require N=12,562 participants.

this will have a clear and observable impact on people's decisions and behavior. Nevertheless, many areas of the vast design space of anthropographics have not been tested, some of which appear promising (see, e.g., discussions in [42]), and it is crucial to test them to advance knowledge.

It is also possible that different ways of measuring prosociality (e.g., observing donations in actual charity campaigns) would lead to different results, perhaps revealing a clear effect where crowdsourced studies could not. Such studies are also needed. Finally, other ways of using visualization to support humanitarian causes deserve to be considered. For example, it can be useful to study how visualizations can help existing donors better allocate their money to charities, which carries the potential of doing far more good than trying to maximize donations for any single charity [45].

ACKNOWLEDGMENTS

Thanks to Paola Valdivia, Ranjini Aravind, Alexandra Ibarra, Qing Chen, Daniel Gondim, Felipe Vieira, Rafael Feitosa, Marie Destandau, Xiyao Wang, and Mickaël Sereno for helping us pilot-test and debug our experiments. The first author gratefully acknowledges the financial support of the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), grant number 88881.188888/2018-01. This work was partly supported by the Agence Nationale de la Recherche (ANR), grant number ANR-19-CE33-0012.

REFERENCES

[1] Aria Alamalhodaei, Alexandra P. Alberda, and Anna Feigenbaum. 2020. *Humanizing data through 'data comics': An introduction to graphic medicine and graphic social science.* Amsterdam University Press, Amsterdam, 347–366.

[2] Scott Bateman, Regan L Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. 2010. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2573–2582. https://doi.org/10.1145/1753326.1753716

[3] C Daniel Batson, Shannon Early, and Giovanni Salvarani. 1997. Perspective taking: Imagining how another feels versus imaging how you would feel. *Personality and social psychology bulletin* 23, 7 (1997), 751–758.

[4] Shlomo Benartzi and Richard H Thaler. 2001. Naive diversification strategies in defined contribution saving plans. *American economic review* 91, 1 (2001), 79–98.

[5] Ralf Bender and Stefan Lange. 2001. Adjusting for multiple testing—when and how? *Journal of clinical epidemiology* 54, 4 (2001), 343–349.

[6] Alberto Betella and Paul FMJ Verschure. 2016. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one* 11, 2 (02 2016), 1–11. https://doi.org/10.1371/journal.pone.0148037

[7] Paul Bloom. 2017. Empathy and its discontents. *Trends in cognitive sciences* 21, 1 (2017), 24–31.

[8] Rita Borgo, Alfie Abdul-Rahman, Farhan Mohamed, Philip W Grant, Irene Reppa, Luciano Floridi, and Min Chen. 2012. An empirical study on using visual embellishments in visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2759–2768.

[9] Michelle A Borkin, Azalea A Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. 2013. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2306–2315.

[10] Jeremy Boy, Anshul Vikram Pandey, John Emerson, Margaret Satterthwaite, Oded Nov, and Enrico Bertini. 2017. Showing people behind data: Does anthropomorphizing visualizations elicit more empathy for human rights data?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA). Association for Computing Machinery, New York, NY, USA, 5462–5474. https://doi.org/10.1145/3025453.3025512

[11] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[12] Sarah Campbell. 2018. *Feeling Numbers The Rhetoric of Pathos in Visualization.* Ph.D. Dissertation. Northeastern University.

[13] Sarah Campbell and Dietmar Offenhuber. 2019. Feeling numbers: The emotional impact of proximity techniques in visualization. *Information Design Journal* 25, 1 (2019), 71–86. https://doi.org/10.1075/idj.25.1.06cam

[14] Fanny Chevalier, Romain Vuillemot, and Guia Gali. 2013. Using concrete scales: A practical framework for effective visual depiction of complex measures. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2426–2435.

[15] Shauna Concannon, Natasha Rajan, Parthiv Shah, Davy Smith, Marian Ursu, and Jonathan Hook. 2020. Brooke Leave Home: Designing a Personalized Film to Support Public Engagement with Open Data. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA). Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376462

[16] Geoff Cumming. 2014. The new statistics: Why and how. *Psychological science* 25, 1 (2014), 7–29.

[17] Geoff Cumming and Robert Calin-Jageman. 2016. *Introduction to the new statistics: Estimation, open science, and beyond.* Routledge, USA.

[18] Geoff Cumming and Sue Finch. 2005. Inference by eye: confidence intervals and how to read pictures of data. *American psychologist* 60, 2 (2005), 170.

[19] Stephan Dickert, Namika Sagara, and Paul Slovic. 2011. Affective motivations to help others: A two-stage model of donation decisions. *Journal of Behavioral Decision Making* 24, 4 (2011), 361–376.

[20] Pierre Dragicevic. 2016. *Fair Statistical Communication in HCI.* Springer International Publishing, Cham, 291–330. https://doi.org/10.1007/978-3-319-26633-6_13

[21] Arvid Erlandsson, Sigrid Møyner Hohle, Erik Løhre, and Daniel Västfjäll. 2018. The rise and fall of scary numbers: The effect of perceived trends on future estimates, severity ratings, and help-allocations in a cancer context. *Journal of Applied Social Psychology* 48, 11 (2018), 618–633.

[22] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[23] Robert J Fisher. 1993. Social desirability bias and the validity of indirect questioning. *Journal of consumer research* 20, 2 (1993), 303–315.

[24] Andrew Gelman and Hal Stern. 2006. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician* 60, 4 (2006), 328–331.

[25] Alexander Genevsky, Daniel Västfjäll, Paul Slovic, and Brian Knutson. 2013. Neural underpinnings of the identifiable victim effect: Affect shifts preferences for giving. *Journal of Neuroscience* 33, 43 (2013), 17188–17196.

[26] Jennifer L Goetz, Dacher Keltner, and Emiliana Simon-Thomas. 2010. Compassion: an evolutionary analysis and empirical review. *Psychological bulletin* 136, 3 (2010), 351.

[27] Steve Haroz, Robert Kosara, and Steven L Franconeri. 2015. Isotype visualization: Working memory, performance, and engagement with pictographs. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (Seoul, Republic of Korea). ACM, Association for Computing Machinery, New York, NY, USA, 1191–1200. https://doi.org/10.1145/2702123.2702275

[28] Jacob Harris. 2015. Connecting with the dots. https://source.opennews.org/articles/connecting-dots/. [Online; accessed 29-January-2019].

[29] Dandan Huang, Melanie Tory, Bon Adriel Aseniero, Lyn Bartram, Scott Bateman, Sheelagh Carpendale, Anthony Tang, and Robert Woodbury. 2014. Personal visualization and personal visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 21, 3 (2014), 420–433.

[30] Alexander Ivanov, Kurtis Thorvald Danyluk, and Wesley Willett. 2018. Exploration & Anthropomorphism in Immersive Unit Visualizations. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada). ACM, Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3170427.3188544

[31] Keith Jensen. 2016. Prosociality. *Current biology* 26, 16 (2016), R748–R752.

[32] Helen Kennedy and Rosemary Lucy Hill. 2018. The feeling of numbers: Emotions in everyday engagements with data and their visualisation. *Sociology* 52, 4 (2018), 830–848.

[33] Kris N Kirby and Daniel Gerlanc. 2013. BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior research methods* 45, 4 (2013), 905–927.

[34] Robert Kosara and Jock Mackinlay. 2013. Storytelling: The next step for visualization. *Computer* 46, 5 (2013), 44–50.

[35] Martin Krzywinski and Naomi Altman. 2013. Points of significance: error bars.

[36] Steve Lambert. 2016. And what do I do now? Using Data Visualization for Social Change. https://c4aa.org/2016/01/data-visualization-for-what/. [Online; accessed 29-January-2019].

[37] Guowei Li, Monica Taljaard, Edwin R Van den Heuvel, Mitchell AH Levine, Deborah J Cook, George A Wells, Philip J Devereaux, and Lehana Thabane. 2017. An introduction to multiplicity issues in clinical trials: the what, why, when and how. *International journal of epidemiology* 46, 2 (2017), 746–755.

[38] Johannes Liem, Charles Perin, and Jo Wood. 2020. Structure and Empathy in Visual Data Storytelling: Evaluating their Influence on Attitude. *Computer Graphics Forum* 39, 3 (2020), 277–289. https://doi.org/10.1111/cgf.13980

arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13980

[39] Daniel Lim and David DeSteno. 2016. Suffering and compassion: The links among adverse life experiences, empathy, compassion, and prosocial behavior. *Emotion* 16, 2 (2016), 175.

[40] Giorgia Lupi. 2017. Data Humanism, the Revolution will be visualized. https://web.archive.org/web/20190915123530/https://www.printmag.com/information-design/data-humanism-future-of-data-visualization/. [Online; accessed 15-September-2019.

[41] Luiz Morais. 2020. *Designing for More than Efficacy: Investigating the Role of Anthropographics on Compassion*. Ph.D. Dissertation. Universidade Federal de Campina Grande.

[42] Luiz Morais, Yvonne Jansen, Nazareno Andrade, and Pierre Dragicevic. 2020. Showing Data about People: A Design Space of Anthropographics. (2020). (under review).

[43] Luiz Morais, Dandara Sousa, and Nazareno Andrade. 2020. Evaluating a Situated and Physical Anthropographic: An In-the-Wild Study. (2020). https://doi.org/10.13140/RG.2.2.10719.69283

[44] Austin Lee Nichols and Jon K Maner. 2008. The good-subject effect: Investigating participant demand characteristics. *The Journal of general psychology* 135, 2 (2008), 151–166.

[45] Toby Ord. 2019. *The moral imperative toward cost-effectiveness in global health*. Oxford University Press, Oxford, 29–36.

[46] Natasha Pirani, BA Ricker, and MJ Kraak. 2020. Feminist cartography and the United Nations Sustainable Development Goal on gender equality: Emotional responses to three thematic maps. *The Canadian Geographer/Le Géographe canadien* 64, 2 (2020), 184–198.

[47] Lisa Charlotte Rost. 2016. A Data Point Moves into a Bar. https://web.archive.org/web/20200318143617/https://lab.dsst.io/slides/33c3/7999.html. [Online; accessed 18-March-2020].

[48] Edward Segel and Jeffrey Heer. 2010. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics* 16, 6 (2010), 1139–1148.

[49] Paul Slovic. 2007. "If I look at the mass I will never act": Psychic numbing and genocide. *Judgment and Decision Making* 2, 2 (2007), 79–95.

[50] Paul Slovic, David Zionts, Andrew K Woods, Ryan Goodman, and Derek Jinks. 2013. *Psychic numbing and mass atrocity*. Princeton University Press NJ, New Jersey, 126–142.

[51] Scott Slovic and Paul Slovic. 2013. Numbers and nerves: Toward an affective apprehension of environmental risk. In *The Feeling of Risk*. Routledge, London, 107–112.

[52] Deborah A Small, George Loewenstein, and Paul Slovic. 2007. Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes* 102, 2 (2007), 143–153.

[53] Daniel Västfjäll, Paul Slovic, Marcus Mayorga, and Ellen Peters. 2014. Compassion fade: Affect and charity are greatest for a single child in need. *PloS one* 9, 6 (2014), e100115.

[54] Jelte M Wicherts, Coosje LS Veldkamp, Hilde EM Augusteijn, Marjan Bakker, Robbie Van Aert, and Marcel ALM Van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology* 7 (2016), 1832.

[55] Mark Ottoni Wilhelm and René Bekkers. 2010. Helping behavior, dispositional empathic concern, and the principle of care. *Social Psychology Quarterly* 73, 1 (2010), 11–32.

[56] Wesley Willett, Yvonne Jansen, and Pierre Dragicevic. 2016. Embedded data representations. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 461–470.

[57] Mushon Zer-Aviv. 2015. DataViz—The UnEmpathetic Art. https://responsibledata.io/2015/10/19/dataviz-the-unempathetic-art/. [Online; accessed 29-January-2019].

[58] Haotian Zhou and Ayelet Fishbach. 2016. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of personality and social psychology* 111, 4 (2016), 493.